D-Hacking

Emily Black
eblack@barnard.edu
Barnard College, Columbia University
New York, New York, USA

Talia Gillis
tbg2117@columbia.edu
Columbia University
New York, New York, USA

Zara Hall
zyh2000@columbia.edu
Columbia University
New York, New York, USA

ABSTRACT

Recent regulatory efforts, including Executive Order 14110 and the AI Bill of Rights, have focused on mitigating discrimination in AI systems through novel and traditional application of anti-discrimination laws. While these initiatives rightly emphasize fairness testing and mitigation, we argue that they pay insufficient attention to *robust* bias measurement and mitigation—and that without doing so, the frameworks cannot effectively achieve the goal of reducing discrimination in deployed AI models. This oversight is particularly concerning given the instability and brittleness of current algorithmic bias mitigation and fairness optimization methods, as highlighted by growing evidence in the algorithmic fairness literature. This instability heightens the risk of what we term *discrimination-hacking* or *d-hacking*, a scenario where, inadvertently or deliberately, the selection of models based on favorable fairness metrics within specific samples could lead to misleading or non-generalizable fairness performance. We term this effect d-hacking because systematically selecting among numerous models to find the least discriminatory one parallels the concept of p-hacking in social science research of selectively reporting outcomes that appear statistically significant resulting in misleading conclusions. In light of these challenges, we argue that AI fairness regulation should not only call for fairness measurement and bias mitigation, but also specify methods to ensure *robust* solutions to discrimination in AI systems. Towards the goal of arguing for robust fairness assessment and bias mitigation in AI regulation, this paper (1) synthesizes evidence of d-hacking in the computer science literature and provides experimental demonstrations of d-hacking, (2) analyzes current legal frameworks to understand the treatment of robust fairness and non-discriminatory behavior, both in recent AI regulation proposals and traditional U.S. discrimination law, and (3) outlines policy recommendations for preventing d-hacking in high-stakes domains.

CCS CONCEPTS

• **Social and professional topics** → **Government technology policy**; • **Computing methodologies** → **Machine learning**; • **Applied computing** → **Law**; • **General and reference** → **Evaluation**; **Measurement**; **Reliability**.

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658928

KEYWORDS

algorithmic fairness, machine learning, AI regulation, disparate impact, algorithmic discrimination, distribution shift, artificial intelligence auditing, anti-discrimination law

ACM Reference Format:
Emily Black, Talia Gillis, and Zara Hall. 2024. D-Hacking. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3630106.3658928

1 INTRODUCTION

There has been increasing focus around defining, measuring, and reporting fairness metrics of algorithms in emergent AI regulation [4, 30]. In addition, the disparate impact doctrine—a cornerstone of anti-discrimination law stating that decision-making systems cannot produce disparate impact across demographic groups unless the system meets a business justification and there is no less discriminatory alternative to achieve that legitimate goal—has been increasingly applied to algorithmic systems [22, 42, 48]. Essentially, measuring and reporting prediction disparities is not only becoming best practice but a keystone of AI regulation.

However, despite these regulatory effort, there is increasing evidence of instability and brittleness of fairness optimization for many applications and debiasing methods. Recent work has pointed to how modeling decisions [14, 18, 35, 37, 57] and distribution shift [6, 8, 13, 29, 39, 56, 75] can lead to meaningful differences in fairness performance, either across different training runs or between training and deployment. Yet, recent AI regulatory efforts do little to acknowledge or address these challenges in their frameworks for measuring, reporting, and reducing AI system bias. Arguably, fairness robustness concerns could be addressed when current high-level regulatory efforts are translated into detailed requirements. However, without explicit recognition of the need for fairness robustness, there is no guarantee that more specific guidelines would address d-hacking. Notably, these high-level guidelines do acknowledge robustness concerns when they pertain to model performance unrelated to fairness, underscoring the importance of high-level recognition in shaping future detailed guidance.

At best, this reality leaves open the possibility that even organizations attempting in good faith to comply with AI regulation may invest time and money into bias mitigation that prove to be ineffective, or even detrimental to anti-discrimination goals, due to their lack of generalizability. At worst, the failure to acknowledge the instability and brittleness of fairness optimization could create opportunities for intentional, ongoing discrimination despite recent regulatory efforts to prevent AI discrimination. This could happen if firms intent on evading discrimination laws manipulate the process by choosing models that appear fair on a specific training set, yet fail to demonstrate generalized fairness across the deployment population.

We term this effect discrimination hacking, or d-hacking, because the search over many models for a model that is least discriminatory can lead to an effect analogous to the concern of p-hacking in social science research. P-hacking involves manipulating data analysis until statistically significant results are obtained, often compromising the integrity and reliability of research findings. Importantly, p-hacking can lead to misleading conclusions by capitalizing on random variations in the data so that findings are not generalizable. Similarly, d-hacking involves searching over models until one exhibits the desirable properties while the true object of interest is the fairness performance out-of-sample.

In light of the potential for d-hacking, intentional or inadvertent, it is crucial to explicitly require robustness analysis for fairness mitigation procedures in any legal requirement for bias mitigation. Towards this goal, in this paper we argue for the incorporation of safeguards against d-hacking in responsible AI regulation. We begin by synthesizing the evidence of the dangers of d-hacking in the computer science literature and provide experimental demonstrations (Section 2). Next, in Section 3 we analyze recent regulatory efforts around AI governance in the U.S. to assess where attention has been paid to the idea of robust AI fairness and where language may be interpreted to require robustness in AI fairness evaluation and mitigation. Although there is some recognition of the risks associated with fragile or unstable fairness measurement and mitigation—indicated by concerns about changes in bias measurements over time, lack of standardization, and lack of replicability—we argue that recent AI regulatory efforts do not adequately prevent d-hacking in practice. Then, in Section 4, we examine traditional U.S. anti-discrimination law to understand how concerns of robustness and generalization have been addressed in more established legal frameworks.

We demonstrate that while robust measures of discrimination have been emphasized, robustness requirements primarily serve as a means to screen discrimination claims. We argue that the similar to the way traditional anti-discriminaiton law has required plaintiffs to demonstrate that measured disparities are not merely product of random variation and particular measurement choices, entities claiming their models are fair should face comparable scrutiny. They must prove that their fairness metrics are robust and reliable. We end, in Section 5, by outlining several policy recommendations for preventing d-hacking in high-stakes domains. Among these, we highlight the need for the same robustness measures applied to ensuring reliable performance to be applied to claims on fairness. We argue that model practitioners should provide an easy-to-use API for systems deployed in a wide range of environments, enabling users to assess deployment data for fairness in their application context. We further recommend practitioners document all analyses conducted and pre-register their intended fairness testing and that model practitioners consider the fairness metrics against hypothetical populations, thereby enhancing fairness and discrimination testing.

2 FAIRNESS INSTABILITY AND D-HACKING

In this section, we present a sample of the evidence for the possibility of d-hacking in the computer science literature. Our aim is to show how a collection of phenomena documented in the machine learning literature— namely, distribution shift, multiplicity, and overfitting—leave open the possibility for practitioners to develop facially fair algorithmic systems that do not perform fairly during deployment, either on accident or on purpose. Instead of presenting a comprehensive survey around problems such as distribution shift or multiplicity as others have done [10, 19], we aim only to expose the reality of these problems and their repercussions to a policy and legal audience, to motivate the need for regulatory attention to ensuring robustly fair systems—which, as we expand in Section 3, is severely lacking.

Several recent works have called attention to the fact that often, models optimized to perform well with regards to fairness at training time do not display similar fairness performance during model deployment. We delve into two sets of reasons why this may be the case—model instability and overfitting, and distribution shift, and how these phenomena relate to d-hacking.

2.1 Model Instability and Overfitting

Recent work has shown that models with almost identical training environments—for example, only different in the random seed set to initialize training [14], a miniscule sampling difference in the training set [14, 23], or even simply a change in the ordering of data points during training [37]—can have noticeable differences in their predictions on individuals. The differences in predictions across models with minute differences in training setup can even aggregate to noticeable differences in group fairness metrics [26, 37]. And, importantly, model instability over small changes to the training environment extends to models trained to satisfy fairness constraints—several works have demonstrated that small differences in model training environment—such as a difference in train-test-split [33, 36] or leave-one-out differences in the training set [47] can have noticeable differences in the fairness they are able to achieve. Indeed, there is evidence to suggest that enforcing fairness on machine learning models can actually increase the model's prediction instability over small perturbations to the training environment [53].

Some have pointed to this instability as a means to increase fairness with little cost to accuracy [15, 19], as often, these models stemming from nearly identical training setups do not differ much with regards to accuracy [19, 53]. While we agree that exploiting model flexibility for decreasing fairness is a good idea, the high variance of fair classifiers presents a risk for d-hacking.

First, this instability in fairness behavior over small perturbations to training setup again leaves the potential for even well-meaning pracitioners to report fairness gains which are simply over-fit to their modeling setup. Much like how the classic bias-variance tradeoff leads to overfitting of high variance models, so too the overfitting of fairness definitions to a particular train-test split could mean the fairness criteria does not generalize to their full deployment distribution. Similarly, a model which is fair on a particular set of hyperparameters may be brittle to online modeling changes made to optimize performance during deployment. Beyond this, practitioners could take advantage of this instability to find a model that looks facially fair but will not perform well on the overall distribution or over modeling changes during deployment. While previous works

Dataset	Model Type	Avg DD (Test)	Avg Accs (Test)	Min. It. LR DD (Test)	Min It. LR Accs (Test)
HMDA	Decision Tree	10.61%	92.76%	12.79%	93.10%
	Logistic Regression	**20.08%**	92.27%	**6.74%**	92.56%
	Random Forest	9.47%	94.03%	10.92%	93.65%
	SVM	9.01%	91.65%	8.40%	91.11%
German Credit	Decision Tree	8.92%	63.45%	6.27%	68.50%
	Logistic Regression	**9.22%**	73.05%	**0.80%**	69.50%
	Random Forest	8.59%	71.80%	3.59%	73.00%
	SVM	8.13%	73.00%	5.01%	74.50%

Table 1: Left two columns: average demographic disparity (DD) and accuracy for different models trained to reduce demographic disparity, evaluated on the model's test set over ten different train/test splits. Right two columns: the demographic disparity and accuracy for the iteration where the model which is worst on average reached the lowest unfairness over the ten runs. The top four rows indicate results for the HMDA dataset, and the bottom four for the German Credit dataset.

have described adversarial methods to create facially fair model *explanations* [7, 9], to the authors knowledge there has been little work around selecting models purposefully to appear fair based on certain metrics or tests but to be unfair during deployment. To illustrate our point, we give a brief experimental demonstration of the possibility for D-hacking as a result of overfitting.

Experimental Demonstration. In these experiments, the protected attribute over which we endeavor to achieve equity is gender, and the metric of fairness we consider is demographic parity, (aiming for equal selection rate, i.e. equal rate of positive predictions, across gender). We test the demographic parity difference of four different classification models (Decision Tree, Logistic Regression, Random Forest, SVM) trained to reduce their demographic disparity using the FairLearn [12] package. We train ten of each of these types of models by creating ten different different random train-test splits of the dataset, and then calculate the accuracy and disparity on the test set for each individual model on the given train-test split and then also the average accuracy and disparity of each model type on average over the ten trials. We perform these experiments over two datasets: the HMDA dataset from Boston [43], and the German Credit Dataset [46]. Experimental details such as the size of the datasets, random seed, and other details can be found in the Appendix.

We demonstrate that which model appears the fairest changes depending on the train-test split, and the fairness can change by an order of magnitude between trials. This instability has been pointed out previously in the literature, as noted in the above paragraphs. However, what we point out here is that this instability can be used for d-hacking: in other words, evidence can be presented to suggest a model is the fairest when it is in fact not.

We present our demonstration in Table 1. In the HMDA dataset, we see that the Logistic Regression model is least fair model when considering average performance over ten random train-test splits, by a factor of two: it displays approximately 20% demographic disparity as opposed to approximately 10% for the other models. However, over the ten train/test splits, the logistic regression model displays an *minimum* demographic disparity of 6.74%—which, when considering only that specific train/test split, is in fact the fairest model. Similarly, on the German Credit Dataset, the Logistic Regression model is again the least fair model when considering average performance over ten random train-test splits. However, on the iteration where it displays the minimum unfairness, the Logistic Regression model appears to be the best choice with respect to fairness—it achieves a demographic disparity almost an order of magnitude lower than the rest of the models (approximately 0.8%). We provide further results on deep models, which have been shown to be particularly unstable in terms of their predictions [14] and fairness [37] across small changes to their training setup, in Appendix A.3.

At best, an unknowledgable practitioner may pick a more discriminatory model, such as the Logisitic Regression model in the case of HMDA, because they did not cross-validate their fairness results over several train-test splits. In situations where there are ulterior motives, however— if a practitioner or company does not want to invest sufficient resources into a more comprehensive search for a less discriminatory model, or if the practitioner or company does not want to sacrifice any accuracy for a more fair model— this instability can be easily harnessed to suggest larger fairness gains than are likely to generalize. In this case, by reporting results from a particular train/test split for fairness, a practitioner can inflate perceived fairness gains while putting in little effort to find a model whose fairness behavior generalizes to the overall population.

Mitigation Techniques. There are some mitigation techniques proposed for this type of model instability— many works have suggested ensembling techniques over a variety of training environments (i.e. using the average response from a large group of models which slightly different training environments) [16, 54]. While these works were aimed at stabilizing prediction behavior generally and not specifically at fairness, they may be able to be extended, or still have a positive effect. Other works have paid more attention to the problem of instability over the samples of the training set in particular, and have developed training techniques which increase robustness over different selections of the training set [27, 33]. Another set of solutions point to creating models whose fairness behavior is robust to a larger set of changes in their treatment distribution, but we cover those works in more detail in the next section.

Dataset	Model Type	D1 Accs (Test)	D1 DD (Test)	D2 Accs	D2 DD
HMDA	Decision Tree	88.98%	6.81%	90.09%	12.07%
	Logistic Regression	81.89%	8.69%	91.82%	30.63%
	Random Forest	91.34%	**1.02**%	94.65%	5.94%
	SVM	91.34%	1.12%	92.92%	**2.68**%
ACS Income	Decision Tree	73.95%	7.83%	70.01%	10.22%
	Logistic Regression	78.97%	3.52%	77.86%	**9.66**%
	Random Forest	80.29%	8.81%	77.12%	12.74%
	SVM	78.54%	**3.26**%	76.27%	11.48%

Table 2: Left two columns: demographic disparity and accuracy for different models trained to reduce demographic disparity trained and evaluated on one distribution (D1). Right two columns: the demographic disparity and accuracy for the same models tested on a shifted distribution (D2). The top four rows indicate results for the HMDA dataset, which was split between two areas in Boston, and the bottom four for the ACSIncome dataset [28].

2.2 Distribution Shift

Another reason why models performing well with regards to fairness or other properties during training and development do not go on to perform well during deployment is *distribution shift.* Distribution shift is a phenomenon where the model's treatment population differs from the dataset that it was trained on. Distribution shift can have different components: the rates of different demographic groups can change from training to deployment data, often referred to as demographic shift; the rate of the predicted label (e.g. default rates in the credit setting) might change in the training and deployment distributions, often called label shift; the prevalence of certain input features in the dataset might change, often called covariate shift; and finally, the actual underlying relationship between the input features and the label might change, so that certain features are no longer predictive—often called concept drift.

While it may seem like an oversight to train a model on a distribution that is different than its eventual treatment population, distribution shift is extremely commonplace in the deployment of AI systems [70]. Any of, or a combination of, these scenarios can happen if a model is trained on data from e.g. a different geographical area than where it will be deployed, as is common in many fairness-critical areas such as pre-trial risk assessment [63]; or a model may be trained on data with a temporal difference from its treatment population, which can happen naturally if a model is in use for many years without retraining; or, improvements in data collection over time might result in distribution shift. Some scholars have shown that releasing models meant to positively benefit society— such as healthcare systems meant to reduce hospitalization—actually change the underlying distribution of their treatment population over time and become unfair as a result of this self-induced distribution shift [59]. Several recent papers provide a more in-depth technical breakdown of the various kinds of distribution shift for interested readers [10], here we focus on highlighting certain results which are relevant to the possibility of d-hacking.

Ding et al. [28] provide a comprehensive display of the inconsistency of fairness results across distribution shifts by creating a variety of datasets varying over geography and time from US Census data, and testing performance of fairness intervention across when trained on one subset and then tested over various changes in time and location. They find both that the "effect size of different [fairness] interventions varies greatly" across subsets of the data (e.g. one state to another), and that "training on one state and testing on another generally leads to unpredictable results. Accuracy and fairness criteria could change in either direction" [28]. Giguere et al. [39] show that common fairness enforcing packages such as FairLearn, and even methods which have been developed to be sensitive to distribution shift, are largely unable to reproduce training-time fairness behavior after a distribution shift. This lack of robustness in fairness behavior over distribution shifts that are incredibly common in machine learning deployments [70] leaves open the possibility that even practitioners trying to enforce fairness on their systems may accidentally over-fit to their particular sample. Even worse, practitioners can take advantage of this lack of robustness to find a fair training method and training distribution to create a model which appears to lead to good fairness results on particular deployment domain, but does not generalize well to the full set of deployments, or distribution shifts.

Experimental Demonstration. In this set of experiments, we show how the phenomenon of distribution shift can also unwittingly or intentionally end in a suboptimally fair model being selected. We train and test four different models on one segment of a distribution– in the case of HMDA [43], we train and test the model on rows from the Boston Suffolk county (D1), and simulate deployment behavior on more rural region in the greater Boston area (D2), whereas for the ACS Income dataset [28], we train and test on data from California (D1), and then simulate deployment in Tennessee (D2). As we can see in both cases, the model selected to be the most fair in the original distribution is *not* the fairest model for the deployment distribution. This leaves open the possibility that a suboptimally fair model is chosen for deployment based on a (potentially even mildly, as in the case of the Boston data) non-representative test set. This phenomenon has been much more extensively documented in, e.g., Ding et al. [28], and simply provide a brief illustration here. Again, we present this demonstration here to point out that distribution shift can be leveraged for d-hacking—i.e., distribution shift can be exploited to deploy suboptimially fair models in practice. Companies that know their deployment populations will be varied should collect data and test the disparity of their model on as much relevant data—however, if a company is unwilling to invest resources to do so, they may simply show that their model is fair on available data, leaving open the possibility of d-hacking. If regulation around discrimination testing is not sufficiently strict—e.g. does not discourage behavior such as incomplete coverage of the deployment distribution in the fairness test data—companies could be compliant while still being discriminatory. Regulation around bias testing and mitigation should thus be designed with d-hacking prevention in mind to effectively stop the deployment of biased systems.

Mitigation Techniques. Thankfully, the problem of fairness generalizability across data shifts has lead to a growing field of work aimed at creating techniques which guarantee fairness across multiple domains [8, 13, 29, 39, 56, 75]. There are certainly practical hurdles to several of these methods—for example, some methods require a complete causal model of the data distribution in order to ensure generalizability of fairness behavior over distribution shift [76]; others require at least some amount of data from the target distribution [25], and others still can lead to unnecessarily large performance drops due to protecting against unfairness over an extremely wide set of possible distributions [80]. Regulatory attention to the problem of instability in fairness mitigation may help push for much needed further development and stress-testing of robustly fair learning paradigms, to help understand which methods might be best to use in practice under various contexts. However, as we expand upon in Section 3, the regulatory frameworks growing around fairness testing and mitigation pay inadequate attention to the *robustness and stability* of reported fairness metrics and mitigation strategies.

3 FAIRNESS AND ROBUSTNESS IN EMERGING AI REGULATION

In recent years, there have been regulatory efforts to establish legal guidance and frameworks for the regulation of AI. A primary focus of these efforts is addressing issues related to fairness and discrimination. This section begins by discussing recent initiatives within the U.S. aimed at establishing a regulatory framework for AI, and specifically their focus regarding fairness and discrimination. We then consider the limited ways in which some of these efforts highlight robustness-related concerns and consider whether they adequately address the risks discussed in Section 2.

We draw three conclusions from our analysis. First, a fundamental aspect of current regulatory efforts is the requirement that models be fair and non-discriminatory, suggesting that robust measurement, testing, and mitigation of bias should be mandated to achieve these stated goals. Second, despite the substantial evidence of d-hacking risks discussed in Section 2, current regulatory efforts largely overlook the robustness of discrimination and fairness measurement. Although some regulatory efforts acknowledge robustness-related concerns, these instances only partially recognize the risks of d-hacking and the necessary approaches to address them. Third, while some may claim that current regulatory efforts are too high-level to recognize the need for fairness robustness, we argue that without some acknowledgement, there is no guarantee that more specific guidelines stemming from these efforts would address d-hacking. Notably, these high-level guidelines do acknowledge robustness concerns when they pertain to model performance unrelated to fairness, underscoring the importance of high-level recognition in shaping future detailed guidance.

3.1 Emerging regulation of AI

Although significant academic attention has already been given to the possibility that algorithmic predictions may raise fairness and equity concerns, these concerns have only recently entered legal and policy discourse. In the U.S., the development of federal regulatory frameworks governing the use of AI both by government agencies and private market participants is still in its infancy. The current approach is to provide overarching guidance through several initiatives of the White House and other bodies, while agencies are primarily responsible for the implementation of the regulatory framework in their respective areas of responsibility.

In this section we provide brief descriptions of some of the most high-profile attempts to regulate AI and highlight the way they emphasize fairness and discrimination. In the next section, we discuss how these initiatives address fairness robustness, if at all.

Blueprint for an AI Bill of Rights. The White House's *Blueprint for an AI Bill of Rights* [4], circulated in October 2022, was one the first attempts to highlight the risks associated with AI and lay down principles to secure the public's rights to enjoy the potential benefits from automated systems while safeguarding against the harms. One of the rights in the Blueprint is "Algorithmic Discrimination Protections" [4]. The Blueprint explains that "[a]lgorithmic discrimination occurs when automated systems contribute to unjustified different treatment or impacts disfavoring people based on their race, color, ethnicity, sex . . ., or any other classification protected by law" [4]. Accordingly, it urges AI developers to take a range of precautionary measures before taking their products to market, including proactively assessing equity considerations during the design phase, ensuring that data inputs are representative and robust, and monitoring and mitigating disparities both before and during the product's use [4].

Executive Order 14110. Building on the *Blueprint for an AI Bill of Rights*, the Biden Administration's Executive Order No. 14110 on *Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence* (E.O. 14110) from October 30, 2023 [30] lays down policies and principles for responsible AI. E.O. 14110 expands upon previously articulated fairness concerns in its discussion of its fourth policy area, which relates to equity and civil rights, reiterating the concern that facially neutral algorithmic inputs may exacerbate existing patterns of discrimination [11].[1] Unlike the Blueprint that lists high-level rights, E.O. 14110 acts as a directive for federal agencies to safeguard against the risks of AI. This agency-by-agency approach allows each federal entity to tailor its guidelines to its respective regulatory domain. For example, E.O. 14110 requires that

[1] "Artificial Intelligence systems deployed irresponsibly have reproduced and intensified existing inequities, caused new types of harmful discrimination, and exacerbated online and physical harms." The definition of algorithmic discrimination that appears in the Blueprint and is reiterated in [30] is also included in the Biden Administration's February 2023 Executive Order No. 13985 on *Further Advancing Racial Equity and Support for Underserved Communities Through the Federal Government* 86 Fed. Reg. 7,009 (Feb. 16, 2023), available at https://www.federalregister.gov/documents/2021/01/25/2021-01753/advancing-racial-equity-and-support-for-underserved-communities-through-the-federal-government.

the Department of Justice furnish a report on the use of AI in the criminal justice system, that the Department of Human Services and Homeland Security publish a plan addressing states' use of algorithmic systems in distributing public benefits, and that the Department of Labor publish guidance for federal contractors to prevent bias in AI systems used in hiring decisions [30]. These directives reflect a preference of maintaining a broad conception of algorithmic discrimination while delegating to relevant agencies the role of developing substantive guidance in specific contexts.

Office of Management and Budget's Memorandum. Following E.O. 14110, the Office of Management and Budget (OMB) released a Draft Memorandum on *Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence* (OMB Draft Memorandum) to implement E.O. 14110 [62]. The draft provides further guidance on AI governance structures in federal agencies by articulating concrete evaluation and monitoring practices and directs federal agencies to take special precautions when dealing with what it terms "rights-impacting AI." The draft memorandum defines rights-impacting AI as AI "whose output serves as a basis for decision or action that has a legal, material, or similarly significant effect" on an individual's civil rights, equal opportunities, or access to critical resources or services. This definition of rights-impacting AI is supplemented with a list of purposes that carry the presumption of fitting within this category, including law enforcement-related risk assessments, tenant screening and controls, employment decisions, loan-allocation processes, and decisions regarding eligibility for government benefits, among others [62].

National Institute of Standards and Technology's Risk Management Framework. A final example of recent AI regulation initiatives is the National Institute of Standards and Technology's (NIST) *Artificial Intelligence Risk Management Framework* (NIST Framework) from January, 2023 [61]. As a federal agency under the U.S. Department of Commerce, NIST specializes in developing measurement standards and advancing technology, including AI. The NIST Framework offers a voluntary, structured approach for integrating responsible AI practices in the design, development, deployment, and assessment of AI products, services, and systems. Developed through multiple iterations and public feedback, the NIST Framework emphasizes the importance of addressing harmful bias and discrimination in AI, marking it as a key aspect of trustworthy AI. One of the characteristics of trustworthy AI laid out in the framework is the mitigation of bias: "Fairness in AI includes concerns for equality and equity by addressing issues such as harmful bias and discrimination" [61].[2]

These initiatives demonstrate the emerging attempt in the U.S. to embrace the benefits of AI while providing a regulatory structure to mitigate the harms and discriminatory concerns of AI and the need to test and monitor algorithmic decision-making for fairness and discrimination purposes. In the next section, we discuss how despite the prominence of testing for discrimination in these proposals, less attention has been paid to the robustness of this testing, potentially undermining the high-level goals of the initiatives in mitigating AI discrimination concerns.

3.2 Fairness robustness in regulatory initiatives

Recent regulatory efforts emphasize that some level of testing is required to ensure that algorithms are deployed in an equitable and nondiscriminatory manner. In this section, we discuss a number of robustness-related concerns mentioned in these efforts. Although these requirements are likely inadequate to address the risks highlighted in Section 2, they reflect a caution against relying on fairness metrics that may not generalize.

Temporal aspects of fairness testing. Some AI initiatives require both an initial consideration of fairness concerns as well as ongoing testing. These added temporal aspects of fairness testing can reflect the concern of robustness of initial fairness measures. For example, the *Blueprint for an AI Bill of Rights* requires the initial showing of fairness and, post-deployment, there must be "ongoing disparity testing and mitigation, and clear organizational oversight" [4]. The OMB draft provides more specific instructions with respect to ongoing monitoring. For example, the OMB draft requires that agencies consider how "unforeseen circumstances, changes to the system after deployment, or changes to the context of use or associated data" may impact discrimination and fairness concerns. Where sufficient mitigation is not possible, agencies must safely discontinue use of the affected AI functionality [62].

Requirements for the ongoing testing of fairness and discrimination also appear in the guidance of several federal agencies who have sought to implement E.O. 14100 by adapting the scope of algorithmic fairness and discrimination to the context of their respective jurisdictions. The Equal Employment Opportunity Commission (EEOC), responsible for enforcing federal laws that prohibit discrimination in the workplace, encourages employers to ensure that AI selection procedures do not produce disproportionately large negative impacts by "conduct[ing] self-analyses on an ongoing basis" [84]. Similarly, the Federal Trade Commission (FTC) has encouraged that lenders comply with lending laws by basing credit decisions on "data derived from an empirical comparison of sample groups . . . that [] are periodically revalidated by the use of appropriate statistical principles and methodology, and adjusted as necessary to maintain predictive ability" [31].

While these guidelines emphasize the need for fairness consideration post-deployment, this ongoing testing requirement receives relatively little attention relative to the initial showing of fairness in these initiatives. Moreover, there is little attempt to specify what these ongoing obligations entail and whether they address the robustness concerns we have highlighted.

Risks due to Lack of Standardization. One aspect of AI risk management that the NIST Framework highlights, although not exclusively in the fairness setting, is the concern over the lack of standardized measures of responsible AI: "The current lack of consensus on robust and verifiable measurement methods for risk and trustworthiness, and applicability to different AI use cases, is an AI risk measurement challenge... measurement approaches can be oversimplified, gamed, lack critical nuance, become relied upon in unexpected ways, or fail to account for differences in affected groups and contexts" [61]. This statement from NIST underscores some of the concrete concerns raised in Section 2. However, despite recognizing these issues, the framework stops short of explaining

[2] The NIST Framework is intended to be updated with the shifting regulatory environment. For example, in April 2024, NIST released its AI Risk Management Framework for Generative AI.

how to practically implement these standards as legal requirements, and does not directly link the lack of standardization to the concern that fairness and discrimination testing can be manipulated or lead to non-generalizable results.

Replicability of Standards. Beyond standardization, another theme in some initiatives is that the auditing of AI systems should produce replicable results. For instance, E.O. 14110 emphasizes that ensuring AI systems are safe and secure requires "robust, reliable, repeatable, and standardized evaluations of AI systems" [30]. This requirement for repeatable evaluations could be interpreted as a need for the replicability of testing results, which may include ensuring that initial fairness results are consistent in future evaluations. E.O. 14110 provides little detail on how to implement auditing that would allow for such repeatable results. Instead, it delegates the responsibility for robustness to federal agencies through a mandate to prevent discrimination via "robust technical evaluations, careful oversight, engagement with affected communities, and rigorous regulation" [30].

While emerging regulatory frameworks acknowledge some robustness related concerns in the estimation and mitigation of bias in AI systems, they fall short in concretely requiring robust fairness evaluation and mitigation to address the risks identified in Section 2. This lack of detailed consideration for discrimination robustness should not be solely attributed to the high-level nature of these frameworks. This is because the frameworks do highlight robustness concerns in contexts unrelated to discrimination, and even when they become more detailed, they tend to overlook fairness and discrimination robustness. For instance, regulatory frameworks discuss robustness in the context of model performance and safety. E.O. 14110 mandates "robust technical evaluations" of AI systems primarily in terms of system safety, but does not extend this robustness requirement to fairness and discrimination. Similarly, the NIST Framework defines robustness or generalizability as the "ability of a system to maintain its level of performance under a variety of circumstances," yet does not connect this robustness requirement to fairness and discrimination. These instances suggest that while drafters of these frameworks were concerned with AI system performance and safety robustness, they did not emphasize fairness and discrimination robustness. Moreover, even when the frameworks provide more detailed guidance, such as the discussion in the 40-page *Blueprint for an AI Bill of Rights* [4] about fairness auditing, they neither recognize nor address concerns related to d-hacking. Our conclusion is that as agencies consider concrete auditing requirements, fairness robustness should be acknowledged to ensure that more specific guidelines effectively address d-hacking.

4 ATTENTION TO ROBUSTNESS IN ANTI-DISCRIMINATION LAW

Consideration of discrimination robustness measurement has implicitly existed in traditional discrimination law. Discrimination law, particularly the disparate impact doctrine, has long been preoccupied with the robustness of demonstrating and measuring disparities for the purpose of identifying a discriminatory policy. Discrimination claims often rely on the statistical demonstration of disparities, raising concerns about the accuracy and robustness of these measurements. While individual treatment and intentional discrimination cases can focus on specific instances of discriminatory conduct, both disparate impact and systematic disparate treatment claims typically involve aggregate statistical analysis.

This section offers an overview of U.S. discrimination law and the role of statistics, especially in establishing a *prima facie* case of disparate impact.[3] We then explore longstanding debates within traditional discrimination law on the robustness of disparity claims. The traditional focus of robustness was primarily to safeguard against spurious lawsuits, and therefore acted as a way to screen and scrutinize plaintiff claims. We argue that similar to the way traditional anti-discrimination law has required plaintiffs to demonstrate that measured disparities are not merely products of random variation and particular measurement choices, entities claiming their models are fair should face comparable scrutiny. They must prove that their fairness metrics are robust and reliable. While the analysis below discusses a number of robustness tests that have long existed in traditional discrimination law, they are not meant to reflect the full extent of AI robustness testing. Rather, the examples of robustness are meant to demonstrate that robustness testing in discrimination has long-existed, strengthening the case for developing appropriate AI robustness testing to address d-hacking.

U.S. Discrimination Law and Disparate Impact. Discrimination law in the United States comprises a network of federal and state (sometimes even municipal [5]) laws designed to address discrimination in various domains by both private and public actors. At the constitutional level, the Equal Protection Clause of the Fourteenth Amendment prohibits states from denying any person within its jurisdiction equal protection under the law. At the federal level there are a number of discrimination statutes that restrict both government and private actors including the Civil Rights Act of 1964, dealing with employment and fair housing, the Age Discrimination in Employment Act (ADEA), the Americans with Disabilities Act (ADA) and the Equal Credit Opportunity Act (ECOA). In the context of voting, the Voting Rights Act of 1965 aims to end racial discrimination in voting.

In some domains of discrimination law, like employment and fair lending, there is a prohibition on both disparate treatment, dealing with intentional discrimination or a direct conditioning of a decision on a protected characteristic, and disparate impact, addressing facially neutral policies that create unjustifiable disparities on the basis of protected characteristics.

In the landmark Supreme Court decision, *Griggs v. Duke Power Co.* [77] and subsequent case law, courts laid out the basic burden shifting framework of disparate impact. In disparate impact litigation, the plaintiff has the initial burden to demonstrate a causal connection between the defendant's policy and the disproportionate effect on a protected class. Upon establishing a *prima facie* case, the onus shifts to the defendant. The defendant can counter by challenging the plaintiff's statistical evidence or by demonstrating that the disputed practice reflects some business necessity [40]. Lastly, if it is shown that the legitimate business goal of the policy can still be achieved by a practice that is less discriminatory, the

[3] Although not our focus, it is important to note that some disparate treatment cases that make pattern claims rather than claims about an individual action can also rely on statistical evidence. See discussion in [20, 74].

defendant faces liability. This last step is often referred to as the "less discriminatory alternative" and requires that even if a practice is justified, there still must not be an alternative way to achieve that business goal that is less discriminatory [15, 41]. Traditionally, discussions of robustness have centered around the first stage of a disparate impact claim—the *prima facie* showing of disparities.

4.1 Robustness in Disparate Impact Claims

There are several ways in which traditional discrimination law has been concerned with the robustness of the measurement of disparities supporting a discrimination claim. Traditionally, robustness requirements were primarily considered as a way to constrain claims of discrimination by plaintiffs. By ensuring that disparities documented by the plaintiff reflect true differences in treatment of groups and not an artifact of some non-generalizable example, courts were able to prevent spurious claims.

It is important to note that traditional claims of discrimination relied on metrics that differ from many of the current algorithmic fairness metrics that look at predictive accuracy [45, 60]. For example, fair lending often considered differences in the decision rates (such as whether to approve a loan application or not). Thus, the Adverse Impact Ratio (AIR) emerged as a standard way to consider disparities by simply comparing the approval rates across groups. For continuous measures, such as differences in interest rates, the traditional measure has been the Standardized Mean Difference (SMD) which is the difference between the average outcome divided by the standard deviation across groups [34]. The AIR and SMD measures of disparities are related to what is referred to in the algorithmic fairness literature as "statistical parity" in which outcomes for different populations are compared [see e.g., 71, "In the context of credit-decision making, evaluations of fairness tend to focus on statistical parity"]. These measures differ from some of the measures of fairness considered in the algorithmic fairness literature, which often focus on error rate measures of prediction of default that are used for lending decisions.[4] Below we provide several examples of robustness discussions in disparate impact doctrine:

Robustness in comparison group. A case of disparate impact relies on the claim that a protected group has been disproportionately impacted by a practice compared to another baseline group. A key question for this comparison is how to define or identify this baseline group to which the protected group is compared. In case law, this is often referred to as a group who is "similarly situated." As previously pointed out by a commentator: "This raises many complex and important questions in the context of fair lending. For example, whether the comparison group involves credit applicants or potential credit applications. Another important question is around who is similarly situated" [55]. In fair lending and employment discrimination this can raise challenges over whether a relevant comparison group are actual applicants or potential applicants. Similarly, in the context of employment discrimination, comparing the impact of a policy on a protected class with its impact on the general population might not be appropriate if the general population is not representative of the qualified job applicant pool.

Robustness in magnitude of disparities. Courts have typically required that a practice be shown to cause "substantial" or "significant" disparities for a disparate impact claim [86]. One way that this requirement has been interpreted is that a plaintiff needs to demonstrate that the magnitude of disparities are practically meaningful. In the context of employment discrimination, this requirement is often associated with the "four-fifths" rule according to which if the magnitude of the ratio of selection rate of a hiring practice for the protected group and the baseline group is below four-fifths, the hiring practice is discriminatory. The four-fifths rule, also known as the 80% rule, was first laid out in the U.S. Equal Employment Opportunity Commission (EEOC) guidelines [3] and has been adopted by several courts. Courts [78] and commentators [64, 85] have highlighted that this ratio should be regarded as more of a rule of thumb than a definitive test for identifying disparate impact. Nonetheless, the rule demonstrates the importance of magnitude when considering disparities.

Robustness in statistical significance. In addition to robustness in magnitude of disparities, courts have emphasized the requirement that claimed disparities be statistically significant for a showing of discrimination. This dimension of robustness relates to the confidence in the disparities identified rather the significance in terms of their magnitude [81]. The 2009 Supreme Court decisions *Ricci v. DeStefano* [78], for example, explains that a prima facie case of disparate-impact liability requires a "threshold showing of a significant statistical disparity" [1, 82].[5] The reason courts often require the showing of statistical significance is the assumption that even if disparities are demonstrated on a certain population, if that disparity has a moderate probability of occurring by chance there is insufficient evidence of disparate impact [38].

Despite the repeated requirement that disparities demonstrate some level of statistical significance, courts and regulators have consistently avoided providing a specific and generally applicable statistical test. For example, the Supreme Court has stated that: "We have emphasized the useful role that statistical methods can have in Title VII cases, but we have not suggested that any particular number of "standard deviations" can determine whether a plaintiff has made out a prima facie case in the complex area of employment discrimination. Nor has a consensus developed around any alternative mathematical standard. Instead, courts appear generally to have judged the "significance" or "substantiality" of numerical disparities on a case-by-case basis" [86]. Similar language exists in the Department of Housing and Urban Development's 2023 Disparate Impact Rule: "HUD further declines to set statistical standards, including statistical thresholds, to require localized statistics, or note a 'significance' requirement [83].

Robustness in causality. In the Supreme Court decision from 2015, *Inclusive Communities*, the requirement that the plaintiff show a robust causal connection between the disparities and challenged policy was laid down as a key requirement of the first stage of a

[4] There continues to be significant debate over the extent to which error rates should be the metric for discrimination in general (see [45]) and whether they are used in practice by lenders in the context of fair lending. See [34]("Stakeholders report substantial variance in the extent to which alternative fairness metrics such as predictive accuracy by group are being used today")

[5] Note that even the EEOC guidelines that discuss the four-fifth rule still state that smaller differences can be disparate impact if they are statistically significant.

disparate impact claim [79]. According to the court "a statistical disparity must fail if the plaintiff cannot point to defendant's policy or policies causing that disparity," a requirement the court refers as a "robust causality requirement." This robustness requirement can create meaningful hurdles for plaintiffs, especially when it is hard to isolate the impact of any individual policy [32]. The Supreme Court has also discussed how a a robust causality requirement ensures that "[r]acial imbalance . . . does not, without more, establish a prima facie case of disparate impact and thus protects defendants from being held liable for racial disparities they did not create" [2].

These examples of robustness requirements in disparate impact law demonstrate the sensitivity of traditional discrimination law to the inadequacy of point estimates in establishing generalizable conclusions. While we use this analysis to demonstrate the pre-existing notion of robustness in discrimination law, addressing d-hacking shifts the focus of robustness to deployers of AI systems who are seeking to establish that their systems are fair and non-discriminatory. As the focus of AI regulation shifts toward proactive auditing and monitoring of AI for fairness purposes, discrimination robustness requirements should no longer center on the burden on plaintiffs but on firms making affirmative claims about the fairness and discrimination of the tools they utilize. Moreover, the types of robustness testing needed to address d-hacking go beyond the examples above, which do not address robustness concerns related to distribution shift or changes in training data, for example.

5 POLICY IMPLICATIONS AND RECOMMENDATIONS

While a comprehensive discussion of the regulatory framework and guidance needed to address the risks of d-hacking is beyond the scope of this paper, in this section we offer several policy recommendations for preventing d-hacking in high-stakes domains. It is crucial to limit the extent to which deployers can, knowingly or unknowingly, inflate their fairness measurements when reporting on their system's discrimination metrics. While existing regulations give some attention to robust fairness assessments and bias mitigation—highlighting the need for ongoing assessments, standardization of testing metrics, and replicable AI assessments—we argue that more concrete recommendations are necessary. To this end, we provide several non-exhaustive suggestions.

Applying Performance Robustness to Fairness. At a minimum, we suggest that techniques used to determine whether or ensure that a model will perform well with regards to *accuracy or performance* should also be applied to fairness measures. For example, performing cross validation and calculating confidence intervals over fairness performance should be common practice for model fairness measurement. This should be straightforward, as most of the necessary infrastructure for such calculations already exists. Similarly, methods that maintain model accuracy across various training setups or treatment distribution perturbations—such as using ensembles, distributionally robust optimization, or transfer learning—should be adapted for fairness purposes. Sections 2.1 and 2.2 discuss several methods to augment or ensure generalization of fairness performance during deployment. Ideally, these methods would be tested more systematically on real-world AI systems and then integrated into training workflows. Although this approach requires a significant investment, it mirrors the commitment made to maintain accuracy in these systems.

Open Fairness API. For systems where deployment scenarios may be extremely varied—for example vendors which sell a base system to users with many different application areas— vendors should provide a non-expert usable API to test deployment data for fairness. If no labeled data is available from the deployment distribution and labels are needed for the desired fairness definition, companies should provide systems that allow users to easily manually label a small amount of their data for testing.

Documentation of Tests and Pre-specification. We also believe there are lessons that can be learned from our initial analogy to p-hacking and the solutions that have been addressed to mitigate the practice and its harm [67]. For example, deployers could be required to record the various analyses they conducted before selecting the model and dataset they used for auditing. This could reduce the incentives to knowingly game fairness metrics and prevent repeated testing for the purpose of obtaining a specific result. This includes reporting on exploratory analyses separately from analyses that are used to report final metrics. Another strategy to prevent p-hacking, which could be adapted for d-hacking, is the pre-registration of intended analysis. In the case of model deployers, they can register upfront and commit ahead of time to the framework they intend to use for testing for fairness and discrimination, requiring them to specify in advance the procedure they intend to utilize. While pre-registration has important limitations in non-experimental settings, as has been discussed in the context of observational social science studies [49], it nonetheless provides an avenue of communicating fairness testing intentions *ex ante.*

Discrimination Stress Testing. In settings where discrimination law is well-established and there is regulatory oversight, fairness testing could be standardized and extended to include regulator-led testing. "Discrimination stress-testing," proposed in the context of fair lending [43], provides an *ex ante* framework for testing models on a hypothetical set of individuals. Rather than relying on measuring of fairness and discrimination in the validation dataset of the deployer, under discrimination stress-testing the regulator would apply the model to some hypothetical population unknown ahead of time to the model deployer. Similar to bank stress-testing, where the financial institution's health is tested under hypothetical scenarios that are often not known in advance to the bank, with discrimination stress-testing the exact data set used to measure disparities could be kept confidential so that model deployers are limited in their ability to create models that minimize disparity for the specific data set alone. We note that there has been work documenting the downsides of confidential datasets for accuracy testing [69], where large swaths of machine learning models were able to overfit even for an unknown test set. However, steps can be taken to prevent or mitigate such overfitting, such as updating the hidden discrimination test set regularly, or using a group of test sets.

6 RELATED WORK

In addition to the technical work summarized in Section 2 of this paper, there have been several works recommending various approaches to algorithmic bias assessments. While this work centers very specifically on the potential for discrimination-hacking and how current regulation responds to that possibility, our piece aligns with some concerns raised in recent algorithmic auditing literature, namely, concerns and limitations around internal audits [24, 66], calls for assessment and monitoring throughout model deployment [17], and the possibility for model developers and systems users to work together to assess an algorithmic system [73]. While several papers call for fairness testing over time, insinuating some concern for robustness to distribution shift, and also call for adversarial testing [52, 66], suggesting that it is important for a model to perform well in unseen environments, very few papers particularly point to the need for *robust fairness testing*. Our paper contributes to this literature by pointing out the possibility for discrimination-hacking, the insufficient nature of the regulatory response to date, and what steps we might take to prevent it.

This paper is also related to the growing literature on algorithmic impact assessments (AIAs). In recent years, scholars across a variety of disciplines have proposed algorithmic impact assessments (AIAs) as a promising means for evaluating several metrics–including algorithmic fairness and discrimination–in practice. Several of these proposals stress the temporal robustness concern, by requiring that AIAs have "periodic revisiting even after their implementation" [21] and that they be "continuous" and include "ongoing assessment and performance evaluation" [50]. The AI Now Institute proposal goes further by suggesting the AIA be renewed every two years and sets down processes for monitoring [68]. Other commentators have raised other issues related to the robustness of fairness metrics, such as whether they adequately map onto real harm [58], whether they are an appropriate optimization object when considering prospective harm [64], or the importance of independent curation of the testing dataset to avoid selective data sampling [44], and the overall concern that deployer discretion and incentives structures may undermine the effectiveness of AIAs [72]. Our focus goes beyond concerns of temporal robustness by discussing how the many degrees of freedom facing deployers can lead to fairness metrics that do not generalize, whether by design or by inattention to robustness. We highlight the need for future developments in AIA frameworks to concentrate more on addressing the risks of d-hacking and the mitigation of associated harms.

7 CONCLUSION

This paper highlights the issue of discrimination-hacking (d-hacking) in AI systems, where the brittleness of fairness measurement and mitigation can be exploited, intentionally or unintentionally, to comply with responsible AI regulation while still deploying biased systems. We demonstrate that, although historical anti-discrimination laws have considered the robustness of discrimination claims, this consideration has been primarily in the context of screening and limiting such claims. As AI deployers increasingly make representations about the fairness of their systems, either voluntarily or as mandated by law, they should be required to demonstrate the robustness of their measures. Despite the risks associated with d-hacking, current regulatory frameworks fall short in requiring robust bias and discrimination measurements, leaving them vulnerable to d-hacking under the guise of compliance.

8 POSITIONALITY STATEMENT

The authors on this work come from different disciplines, ranging from computer scientists to law professors. While this interdisciplinary team does have expertise over computer science and certain aspects of anti-discrimination law, we do not have expertise all areas that are pertinent to the problem of determining the best course for AI regulation. While this paper focuses primarily on the legal doctrines and institutions in the U.S., we recognize that key developments in AI regulation of fairness and discrimination are happening outside of the U.S., and that other countries have developed discrimination doctrines that can shed light on the issue of robustness. We look forward to, and are excited by the prospect of, collaborating with scholars with expertise outside of the U.S. in the future.

9 RESEARCH ETHICS AND SOCIAL IMPACT STATEMENT

The development of this work did not require any data collection. All datasets used were publicly available, on licenses suitable for public use. Since our experimentation was minimal, the majority of our ethical considerations concern the potential impacts of our paper.

Some potential negative impacts of our work include the possibility that the adoption of our policy recommendations will still lead to a checklist-based approach to algorithmic auditing and bias mitigation that does not account for potentially larger issues around AI system deployment, such as issues of AI functionality [65]. While this may be a risk, we note that our proposal is not the only tool we have to combat inequitable algorithmic systems. We can, and should, use a variety of methods to prevent algorithmic inequity. Further, we hope that despite this risk, the added attention to robust fairness evaluation will bring to light more discriminatory systems that need to be changed than it would let discriminatory systems slip under the radar.

ACKNOWLEDGMENTS

We are deeply grateful to Pauline Kim and Haggai Porat for their insightful comments and invaluable suggestions as well as to our anonymous reviewers. We are also grateful to Sahil Soni for excellent research assistance.

REFERENCES

[1] 1982. Connecticut v. Teal. 457 U.S. 440. Available at: https://supreme.justia.com/cases/federal/us/457/440/.
[2] 1989. Wards Cove Packing Co. v. Atonio. , 642, 653 pages.
[3] 2016. 29 C.F.R. § 1607.4(D). Code of Federal Regulations. Available at: https://www.ecfr.gov/current/title-29/subtitle-B/chapter-XIV/part-1607/section-1607.4.
[4] 2022. Blueprint for an AI bill of rights. https://www.whitehouse.gov/ostp/ai-bill-of-rights/
[5] 2023. Local Law 144 of 2021: Automated Employment Decision Tools (AEDT). New York City Legislation. Effective Date: January 1, 2023. This law requires a bias audit on automated employment decision tools before their use and mandates notification to candidates or employees in the city about the use of such tools. Enforcement begins July 5, 2023.

[6] Ashrya Agrawal, Florian Pfisterer, Bernd Bischl, Francois Buet-Golfouse, Srijan Sood, Jiahao Chen, Sameena Shah, and Sebastian Vollmer. 2020. Debiasing classifiers: is reality at variance with expectation? *arXiv preprint arXiv:2011.02407* (2020).

[7] Ulrich Aïvodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. 2019. Fairwashing: the risk of rationalization. In *International Conference on Machine Learning*. PMLR, 161–170.

[8] Bang An, Zora Che, Mucong Ding, and Furong Huang. 2022. Transferring fairness under distribution shifts via fair consistency regularization. *Advances in Neural Information Processing Systems* 35 (2022), 32582–32597.

[9] Christopher Anders, Plamen Pasliev, Ann-Kathrin Dombrowski, Klaus-Robert M¨uller, and Pan Kessel. 2020. Fairwashing explanations with off-manifold detergent. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, 314–323. https://proceedings.mlr.press/v119/anders20a.html

[10] Ainhize Barrainkua, Paula Gordaliza, Jose A Lozano, and Novi Quadrianto. 2023. Preserving the Fairness Guarantees of Classifiers in Changing Environments: a Survey. *Comput. Surveys* (2023).

[11] Biden Administration. 2023. Further Advancing Racial Equity and Support for Underserved Communities Through the Federal Government. Executive Order No. 13985, 86 Fed. Reg. 7,009. Available at: https://www.federalregister.gov/documents/2021/01/25/2021-01753/advancing-racial-equity-and-support-for-underserved-communities-through-the-federal-government.

[12] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. *Fairlearn: A toolkit for assessing and improving fairness in AI*. Technical Report MSR-TR-2020-32. Microsoft. https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/

[13] Arpita Biswas and Suvam Mukherjee. 2021. Ensuring Fairness under Prior Probability Shifts. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Virtual Event, USA) *(AIES ’21)*. Association for Computing Machinery, New York, NY, USA, 414–424. https://doi.org/10.1145/3461702.3462596

[14] Emily Black and Matt Fredrikson. 2021. Leave-one-out Unfairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 285–295.

[15] Emily Black, John Logan Koepke, Pauline Kim, Solon Barocas, and Mingwei Hsu. 2023. Less Discriminatory Algorithms. *Available at SSRN* (2023).

[16] Emily Black, Klas Leino, and Matt Fredrikson. 2022. Selective Ensembles for Consistent Predictions. *ICLR* (2022).

[17] Emily Black, Rakshit Naidu, Rayid Ghani, Kit T. Rodolfa, Daniel Ho, and Hoda Heidari. 2023. Toward Operationalizing Pipeline-aware ML Fairness: A Research Agenda for Developing Practical Guidelines and Tools. *2023 ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (2023).

[18] Emily Black, Manish Raghavan, and Solon Barocas. 2022. Model Multiplicity: Opportunities, Concerns, and Solutions. In *ACM FAccT 2022*.

[19] EMILY BLACK, MANISH RAGHAVAN, and SOLON BAROCAS. 2022. Model Multiplicity: Opportunities, Concerns, and Solutions. *ACM FAccT* (2022).

[20] Kingsley R Browne. 1993. Statistical proof of discrimination: beyond damned lies. *Wash. L. Rev.* 68 (1993), 477.

[21] Alessandra Calvi and Dimitris Kotzinos. 2023. Enhancing AI fairness through impact assessment in the European Union: a legal and computer science perspective. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1229–1245.

[22] Relman Colfax. 2021. Fair Lending Monitorship of Upstart Network’s Lending Model: Initial Report of the Independent Monitor. https://www.relmanlaw.com/media/cases/1088_Upstart%20Initial%20Report%20-%20Final.pdf.

[23] A Feder Cooper, Solon Barocas, Christopher De Sa, and Siddhartha Sen. 2023. Variance, Self-Consistency, and Arbitrariness in Fair Classification. *arXiv preprint arXiv:2301.11562* (2023).

[24] Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini. 2022. Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1571–1583.

[25] Amanda Coston, Karthikeyan Natesan Ramamurthy, Dennis Wei, Kush R Varshney, Skyler Speakman, Zairah Mustahsan, and Supriyo Chakraborty. 2019. Fair transfer learning with missing protected attributes. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 91–98.

[26] Amanda Coston, Ashesh Rambachan, and Alexandra Chouldechova. 2021. Characterizing Fairness Over the Set of Good Models Under Selective Labels. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 2144–2155. https://proceedings.mlr.press/v139/coston21a.html

[27] Andrew Cotter, Maya Gupta, Heinrich Jiang, Nathan Srebro, Karthik Sridharan, Serena Wang, Blake Woodworth, and Seungil You. 2019. Training Well-Generalizing Classifiers for Fairness Metrics and Other Data-Dependent Constraints. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 1397–1405. https://proceedings.mlr.press/v97/cotter19b.html

[28] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems* 34 (2021), 6478–6490.

[29] Wei Du and Xintao Wu. 2021. Fair and robust classification under sample selection bias. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 2999–3003.

[30] EO14110 [n. d.]. Executive Order 14110, Safe, Secure, and Trustworthy Development of Artificial Intelligence, 88 Fed. Reg. 75191 (Nov. 1 2023).

[31] Federal Trade Commission. 2020. Using Artificial Intelligence and Algorithms. https://www.ftc.gov/business-guidance/blog/2020/04/using-artificial-intelligence-and-algorithms.

[32] Kimberly Ferrari. 2020. The State of Disparate Impact Under the Fair Housing Act: Interpreting Robust Causality After Inclusive Communities. *Journal of Affordable Housing and Community Development Law* (2020).

[33] Julien Ferry, Ulrich Aivodji, Sébastien Gambs, Marie-José Huguet, and Mohamed Siala. 2023. Improving fairness generalization through a sample-robust optimization method. *Machine Learning* 112, 6 (2023), 2131–2192.

[34] FinRegLab. 2023. Explainability and Fairness in Machine Learning for Credit Underwriting. https://finreglab.org/wp-content/uploads/2023/12/FinRegLab_2023-12-07_Research-Report_Explainability-and-Fairness-in-Machine-Learning-for-Credit-Underwriting_Policy-Analysis.pdf. Accessed: January 20, 2024.

[35] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. 2019. A Comparative Study of Fairness-Enhancing Interventions in Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* ’19)*. Association for Computing Machinery, New York, NY, USA, 329–338. https://doi.org/10.1145/3287560.3287589

[36] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*. 329–338.

[37] Prakhar Ganesh, Hongyan Chang, Martin Strobel, and Reza Shokri. 2023. On The Impact of Machine Learning Randomness on Group Fairness. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1789–1800.

[38] Joseph L Gastwirth. 1992. Employment discrimination: A statistician’s look at analysis of disparate impact claims. *Law & Ineq.* 11 (1992), 151.

[39] Stephen Giguere, Blossom Metevier, Yuriy Brun, Philip S. Thomas, Scott Niekum, and Bruno Castro da Silva. 2022. Fairness Guarantees under Demographic Shift. In *International Conference on Learning Representations*. https://openreview.net/forum?id=wbPObLm6ueA

[40] Talia Gillis. 2024. “Price Discrimination” Discrimination. (May 2024). Working Paper.

[41] Talia Gillis, Vitaly Meursault, and Berk Ustan. 2024. Operationalizing the Search for Less Discriminatory Alternatives in Fair Lending. In *ACM FAccT 2024*.

[42] Talia B Gillis. 2021. The input fallacy. *Minn. L. Rev.* 106 (2021), 1175.

[43] Talia B Gillis and Jann L Spiess. 2019. Big data and discrimination. *The University of Chicago Law Review* 86, 2 (2019), 459–488.

[44] Ali Hasan, Shea Brown, Jovana Davidovic, Benjamin Lange, and Mitt Regan. 2022. Algorithmic Bias and Risk Assessments: Lessons from Practice. *Digital Society* 1, 2 (2022), 14.

[45] Deborah Hellman. 2008. *When Is Discrimination Wrong?* Harvard University Press.

[46] Hans Hofmann. 1994. Statlog (German Credit Data). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5NC77.

[47] Lingxiao Huang and Nisheeth Vishnoi. 2019. Stable and fair classification. In *International Conference on Machine Learning*. PMLR, 2879–2890.

[48] Mike Isaac. 2023. Meta Agrees to Alter Ad Technology in Settlement With U.S.

[49] Alan M Jacobs, Colin Elman, John Gerring, and James Mahoney. 2020. Pre-registration and results-free review in observational and qualitative research. *The production of knowledge: Enhancing progress in social science* 2020 (2020), 221–264.

[50] Margot E Kaminski and Gianclaudio Malgieri. 2021. Algorithmic impact assessments under the GDPR: producing multi-layered explanations. *International Data Privacy Law* 11, 2 (2021), 125–144.

[51] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. arXiv:1711.05144 [cs.LG]

[52] Xiaoxuan Liu, Ben Glocker, Melissa M McCradden, Marzyeh Ghassemi, Alastair K Denniston, and Lauren Oakden-Rayner. 2022. The medical algorithmic audit. *The Lancet Digital Health* 4, 5 (2022), e384–e397.

[53] Carol Xuan Long, Hsiang Hsu, Wael Alghamdi, and Flavio Calmon. 2023. Individual Arbitrariness and Group Fairness. In *Thirty-seventh Conference on Neural Information Processing Systems*.

[54] Carol Xuan Long, Hsiang Hsu, Wael Alghamdi, and Flavio P Calmon. 2023. Arbitrariness Lies Beyond the Fairness-Accuracy Frontier. *arXiv preprint*

arXiv:2306.09425 (2023).

[55] Peter E Mahoney. 1998. The End (s) of Disparate Impact: Doctrinal Reconstruction, Fair Housing and Lending Law, and the Anti-Discrimination Principle. *Emory LJ* 47 (1998), 409.

[56] Debmalya Mandal, Samuel Deng, Suman Jana, Jeannette Wing, and Daniel J Hsu. 2020. Ensuring fairness beyond the training data. *Advances in neural information processing systems* 33 (2020), 18445–18456.

[57] Charles T. Marx, Flávio du Pin Calmon, and Berk Ustun. 2019. Predictive Multiplicity in Classification. *CoRR* abs/1909.06677 (2019). http://arxiv.org/abs/1909.06677

[58] Jacob Metcalf, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. 2021. Algorithmic impact assessments and accountability: The co-construction of impacts. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 735–746.

[59] Alan Mishler and Niccolò Dalmasso. 2022. Fair when trained, unfair when deployed: Observable fairness measures are unstable in performative prediction settings. *arXiv preprint arXiv:2202.05049* (2022).

[60] Thomas B Nachbar. 2020. Algorithmic fairness, algorithmic discrimination. *Fla. St. UL Rev.* 48 (2020), 509.

[61] National Institute of Standards and Technology. 2023. Artificial Intelligence Risk Management Framework. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

[62] Office of Management and Budget. 2023. Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence. https://www.whitehouse.gov/wp-content/uploads/2023/11/AI-in-Government-Memo-draft-for-public-review.pdf. OMB Draft Memorandum.

[63] Pretrial Risk Assessment. [n. d.]. Where are PRAI Being Used? https://pretrialrisk.com/national-landscape/where-are-prai-being-used/. Accessed: YYYY-MM-DD.

[64] Manish Raghavan and Pauline Kim. 2023. Limitations of the'Four-Fifths Rule'and Statistical Parity Tests for Measuring Fairness. (2023).

[65] Inioluwa Deborah Raji, I Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The fallacy of AI functionality. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 959–972.

[66] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 33–44.

[67] Dorota Reis and Malte Friese. 2022. The Myriad Forms of p-Hacking. In *Avoiding Questionable Research Practices in Applied Psychology*. Springer, 101–121.

[68] Dillon Reisman, Jason Schultz, Kate Crawford, and Meredith Whittaker. 2018. Algorithm Impact Assessment: A Practical Framework for Public Agency Accountability. https://ainowinstitute.org/publication/algorithmic-impact-assessments-report-2. Accessed: 2024-01-21.

[69] Rebecca Roelofs, Vaishaal Shankar, Benjamin Recht, Sara Fridovich-Keil, Moritz Hardt, John Miller, and Ludwig Schmidt. 2019. A meta-analysis of overfitting in machine learning. *Advances in Neural Information Processing Systems* 32 (2019).

[70] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[71] Nicholas Schmidt and Bryce Stephens. 2019. An introduction to artificial intelligence and solutions to the problems of algorithmic discrimination. *arXiv preprint arXiv:1911.05755* (2019).

[72] Andrew D Selbst. 2021. AN INSTITUTIONAL VIEW OF ALGORITHMIC IMPACT. *Harvard Journal of Law & Technology* 35, 1 (2021).

[73] Hong Shen, Alicia DeVos, Motahhare Eslami, and Kenneth Holstein. 2021. Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–29.

[74] Elaine W Shoben. 1983. The use of statistics to prove intentional employment discrimination. *Law and Contemporary Problems* 46, 4 (1983), 221–245.

[75] Harvineet Singh, Rina Singh, Vishwali Mhasawade, and Rumi Chunara. 2021. Fairness violations and mitigation under covariate shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 3–13.

[76] Shubham Singh, Bhuvni Shah, Chris Kanich, and Ian A Kash. 2022. Fair decision-making for food inspections. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–11.

[77] Supreme Court of the United States. 1971. *Griggs v. Duke Power Co.* 401 U.S. 424.

[78] Supreme Court of the United States. 2009. *Ricci v. DeStefano.* 557 U.S. 557.

[79] Supreme Court of the United States. 2015. Texas Department of Housing and Community Affairs v. Inclusive Communities Project, Inc. , 519, 542 pages. Inclusive Communities Project, Inc..

[80] Bahar Taskesen, Viet Anh Nguyen, Daniel Kuhn, and Jose Blanchet. 2020. A distributionally robust approach to fair classification. *arXiv preprint arXiv:2007.09530* (2020).

[81] Kevin Tobia. 2017. Disparate statistics. *The Yale Law Journal* (2017), 2382–2420.

[82] ugesp 1978. Uniform Guidelines on Employment Selection Procedures. 29 C.F.R. § 1607.

[83] U.S. Department of Housing and Urban Development (HUD). 2023. Reinstatement of HUD's Discriminatory Effects Standard. Federal Register. , 19450-19500 pages. Final rule, Document Number: 2023-05836.

[84] U.S. Equal Employment Opportunity Commission. 2023. Select Issues for Assessing Adverse Impact in Software Algorithms and Artificial Intelligence. https://www.eeoc.gov/laws/guidance/select-issues-assessing-adverse-impact-software-algorithms-and-artificial.

[85] Elizabeth Anne Watkins, Michael McKenna, and Jiahao Chen. 2022. The four-fifths rule is not disparate impact: a woeful tale of epistemic trespassing in algorithmic fairness. *arXiv preprint arXiv:2202.09519* (2022).

[86] watsonbank 1988. Watson v. Fort Worth Bank and Trust.

A APPENDIX

A.1 Experimental Details.

For all experiments, we used the sklearn and Fairlearn [12] package to train SVM, Logistic Regression, Random Forest, and Decision Tree classification models with their default parameters on the relevant datasets. We train the models to enforce demographic parity using the gridsearch method, with a gridsize of 10, using the default parameters for the demographic parity constraint. For the cross-validation experiments, we used the German Credit Dataset [46] (n=1000), and the Boston HMDA dataset as gathered and cleaned in [43] (n=2754). For the distribution shift experiment, we used the same HMDA dataset, and also the ACSIncome dataset from California (n=195665) and Tennessee (n=34003), both from 2018 [28]. The German Credit dataset was not well set up for a distribution shift experiment (there were no clear sub-distributions within the data), so we substituted in the ACSIncome dataset.

For the cross validation experiments, we set a random seed of zero at the top of a Jupyter Notebook. We then trained ten different SVM, Logistic Regression, Random Forest, and Decision Tree models (with default parameters for both sklearn and Fairlearn, as explained above) over ten different random 80/20 train-test splits of the data. In our results, we report the average test accuracy and fairness over these train/test splits for all four models, as well as the test accuracy and fairness for the iteration where the least fair model on average performs the best with respect to fairness.

For the distribution shift experiments, we would train and test each of the four models on one distribution (D1) and test its "deployment" behavior on a dataset with some distribution shift. For the HMDA [43] dataset, we train and test the model on rows from the Boston Suffolk county (D1), and simulate deployment behavior on more rural region in the greater Boston area (D2), whereas for the ACS Income dataset [28], we train and test on data from California (D1), and then simulate deployment in Tennessee (D2). We trained and tested on a random split of the data (random seed of 42), which was 10% of the California data for training and 5% for testing, and 10% of the Tennessee data for the deployment set.

A.2 Additional Results on Fairness Gerrymandering.

One related work which we draw particular attention to distinguish from is Kearns at al [51]—this work addresses a different failure mode of fairness enforcement. Namely, the main concern of Kearns et al. [51] is that a classifier that appears fair across certain groups (black/white and men/women) may actually be unfair across unspecified subgroups (black men/white women). In contrast, our main concern is that a classifier may appear to be fair

Model Type	Average DD (Test)	Iteration 1 DD (Test)
Decision Tree	3.22%	6.37%
Logistic Regression	**9.85%**	**0.85%**
Random Forest	8.25%	3.70%
SVM	4.55%	2.88%

Table 3: Left column: Average demographic disparity (DD) for different models trained to reduce demographic disparity, specifically with the fictitious play algorithm guaranteeing subgroup fairness, implemented in the GerryFair package by Kearns et al., evaluated on the model's test set over five different train/test splits. Right column: the demographic disparity for each model for iteration 1. Note that the Logistic Regression model has the lowest demographic disparity in iteration 1, but the highest on average, leaving open the possibility of d-hacking. The results are over the German Credit dataset.

in development and perform unfairly at deployment (e.g., due to overfitting). To highlight this difference, we performed additional experiments to demonstrate the potential for d-hacking even when using fairness training methods designed to mitigate fairness gerrymandering [51]. Similar to Section 2.1 of the main paper, we created 5 random training sets and compared four model types (Decision Tree, Logistic Regression, Random Forest, SVM) after training them to enforce subgroup fairness using the GerryFair package [51]. While the Logistic Regression model has the lowest DD of 0.83% on Iteration 1 (out of a 0.83-6.37% range across models), it actually has the highest average DD of 9.85% across all training sets (out of a 3.22-9.85% range). This demonstrates how models may exhibit low unfairness on a particular training set due to overfitting, but fail to generalize that fairness. Without robustness testing, companies could take advantage of this to artificially report low unfairness. Table 3 shows the results on the German Credit dataset. The left column shows the average demographic disparity (DD) across the 5 training sets for each model type. The right column shows the DD for each model type on just the first training set (Iteration 1). Our results show d-hacking can still occur even using such fair training methods.

A.3 Deep Model Results

While we focus in the main paper on simpler models, here we provide evidence that deep models are vulnerable to d-hacking as well— as is consistent with related work demonstrating that more complex models exhibit more prediction instability over training environment perturbations than simpler models [14]. Our results are presented in Table 4.

Our experimental methodology is the same as laid out in the main paper: for each model type (in this case, different architechture), we train ten different models over ten random train-test splits, and then calculate the accuracy and demographic disparity for each individual model as well as the average across all ten models. For all experiments, we used the sklearn and Fairlearn package to train four different Neural Network classification models with varying architectures. We train the models to enforce demographic parity using the GridSearch method, with a grid size of 10, using the default parameters for the demographic parity constraint. All models use ReLU activation in the hidden layers, sigmoid activation in the output layer, and are compiled with the Adam optimizer, binary cross-entropy loss, and accuracy as the evaluation metric. The four Neural Network architectures are defined as follows:

- 'Neural Network 1': One hidden layer with 100 units.
- 'Neural Network 2': Two hidden layers with 100 and 50 units, respectively.
- 'Neural Network 3': Three hidden layers with 100, 50, and 25 units, respectively.
- 'Neural Network 4': Three hidden layers with 100 units each.

We set a random seed of zero at the top of the Jupyter notebook for reproducibility. We then trained the four Neural Network models over ten different random 80/20 train-test splits of the data. In our results, we report the average test accuracy and fairness over these train/test splits for all four models, as well as the test accuracy and fairness for the iteration where the least fair model on average performs the best with respect to fairness.

We see that, as in the results for the main paper, individual training iterations can have deceptively low disparity (i.e., appear more fair), while actually performing quite poorly in terms of fairness on average over ten train-test splits. In particular, this can lead to choosing a model which appears to be the fairest on one iteration but is actually the least fair over ten train test splits— for example, on the HMDA data in Table 4, note that there exists an iteration where Neural Network 4 has the lowest disparity, but its average disparity is the largest, and over twice that of the lowest average disparity model.

Dataset	Model Type	Avg DD (Test)	Avg Accs (Test)	Min. It. LR DD (Test)	Min It. LR Accs (Test)
HMDA	Neural Network 1	6.97%	91.14%	6.68%	91.65%
	Neural Network 2	5.18%	91.32%	4.31%	92.56%
	Neural Network 3	8.05%	90.93%	0.86%	91.47%
	Neural Network 4	**10.78%**	91.83%	**0.69%**	91.83%
German Credit	Neural Network 1	8.60%	72.40%	11.81%	70.00%
	Neural Network 2	**9.39**%	71.05%	**0.40**%	67.50%
	Neural Network 3	9.27%	70.45%	1.09%	68.00%
	Neural Network 4	8.78%	71.55%	6.74%	68.50%

Table 4: Left two columns: average demographic disparity (DD) and accuracy for different models trained to reduce demographic disparity, evaluated on the model's test set over ten different train/test splits. Right two columns: the demographic disparity and accuracy for the iteration where the model which is worst on average reached the lowest unfairness over the ten runs. The top four rows indicate results for the HMDA dataset, and the bottom four for the German Credit dataset.